remark that sensitivities are continuous functions of b. Thus, every neighborhood of a locally identifiable b_0 contains an open region $G \subset \Omega$ such that the system is "sensitivity" identifiable for all $b \in G$, although the system may be "sensitivity" nonidentifiable at b_0.

References

[1] M. S. Grewal and K. Glover, *IEEE Trans. Automat. Contr.*, vol. AC-21, pp. 833–837, Dec. 1976.

[2] J. Dieudonné, *Foundations of Modern Analysis*. New York: Academic, 1960.

[3] C. T. Chen, *SIAM J. Appl. Math.*, vol. 15, pp. 1272–1274, Sept. 1967.

Maximum Likelihood Parameter Estimation for Linear Systems with Singular Observations

H. KWAKERNAAK

***Abstract*—It is shown that maximum likelihood estimation of unknown parameters of a linear system with singular observations in general results in the maximization of a likelihood function subject to equality constraints.**

I. Introduction

Consider the linear, time-invariant, discrete-time system

$$x(k+1)=A(k;\theta)x(k)+B(k;\theta)u(k)+v(k),$$
$$y(k)=C(k;\theta)x(k)+w(k), \tag{1}$$

$k=0,1,\cdots$, where $v(k)$ and $w(k)$, $k=0,1,\cdots$ are sequences of independent random Gaussian vectors with zero expectations and $\operatorname{var}[v(k)]=V_{11}(k;\theta)$, $\operatorname{cov}[v(k),w(k)]=V_{12}(k;\theta)$, $\operatorname{var}[w(k)]=V_{22}(k;\theta)$, $x(0)$ is a given Gaussian random vector, independent of the sequences v and w, with mean and variance dependent upon θ, and where, finally, θ is an unknown parameter. The parameter θ is to be estimated from an observed record of the output variable $y(k)$, $k=0,1,\cdots,N$, using the fact that the inputs $u(k)$, $k=0,1,\cdots,N-1$, that were applied are known.

The maximum likelihood approach to this estimation problem is well-documented in the literature (see, e.g., Sage and Melsa [1], Goodwin and Payne [2]). It is generally assumed that the observation noise variance matrices $V_{22}(k;\theta)$, $k=0,1,\cdots$ are nonsingular for all θ in the region of interest. We shall study the case that for some or all k, these variance matrices are singular, a situation that may easily arise. Under these circumstances, it is not clear that the likelihood function, which is taken as the joint probability density $p_{y(0),y(1),\cdots,y(N)}(y_0,y_1,\cdots,y_N)$ of $y(0),y(1),\cdots,y(N)$, at all exists. If it does not, a different notion of maximum likelihood estimation is required.

II. Maximum Likelihood Estimation for Singular Gaussian Random Vectors

We shall first see how the idea of maximum likelihood estimation may be modified in the case of a singular Gaussian probability distribution. Let z be a Gaussian random vector, with $Ez=m$ and $\operatorname{var}(z)=\Sigma$. Both m and Σ depend on an unknown parameter θ. Allowing Σ to be singular, it is always possible to find an orthogonal transformation matrix T such that

$$T=\begin{pmatrix}T_1\\T_2\end{pmatrix},\quad T\Sigma T'=\begin{pmatrix}\Delta & 0\\0 & 0\end{pmatrix},\quad \Sigma=(T_1',T_2')\begin{pmatrix}\Delta & 0\\0 & 0\end{pmatrix}\begin{pmatrix}T_1\\T_2\end{pmatrix}$$

where the prime denotes the transpose and where Δ is diagonal and positive definite. It immediately follows that the random variable T_1z is Gaussian, with mean $E(T_1z)=T_1m$, and nonsingular variance $\operatorname{var}(T_1z)=\Delta$. Furthermore, the vector T_2z has zero variance, and hence, with probability one is given by $T_2z=T_2m$. Here, besides m and Σ, T_1 and T_2 also depend on θ.

Under these conditions, the idea of maximum likelihood estimation obviously results in maximizing with respect to θ the density of the nonsingular random variable T_1z, which is given by

$$\frac{1}{(2\pi)^{k/2}[\det(\Delta)]^{1/2}}\exp\left[-\frac{1}{2}(T_1\zeta-T_1m)'\Delta^{-1}(T_1\zeta-T_1m)\right], \tag{2}$$

subject to the equality constraint $T_2\zeta=T_2m$. Here ζ is the observed realization of the random variable z, while $k=\dim(T_1z)$. Thus, instead of dealing with the unconstrained maximization of a likelihood function, we are now facing a maximization with an equality constraint.

The problem may be reformulated in terms of generalized inverses. Defining Σ^+ as the generalized inverse of Σ (see, e.g., Noble [3]), we have $\Sigma^+=T_1'\Delta^{-1}T_1$, so that the likelihood functional (2) can be rewritten as

$$\frac{1}{(2\pi)^{k/2}[\det(\Delta)]^{1/2}}\exp\left[-\frac{1}{2}(\zeta-m)'\Sigma^+(\zeta-m)\right].$$

Since T_2 has full rank, $T_2\zeta=T_2m$ is equivalent to $T_2'T_2\zeta=T_2'T_2m$. As $\Sigma\Sigma^+=T_1'T_1=I-T_2'T_2$, the equality constraint can be rewritten as

$$(I-\Sigma\Sigma^+)(\zeta-m)=0.$$

It is useful to state another formulation of the problem, avoiding generalized inverses. Define $\Sigma^+(\zeta-m)=x$. Then the condition $(I-\Sigma\Sigma^+)(\zeta-m)=0$ is equivalent to $\Sigma x=\zeta-m$. Consequently, $(\zeta-m)'\Sigma^+\cdot(\zeta-m)=(\zeta-m)'x=x'\Sigma x$. Hence, the maximum likelihood method amounts to maximizing

$$\frac{1}{(2\pi)^{k/2}[\det(\Delta)]^{1/2}}\exp\left(-\frac{1}{2}x'\Sigma x\right)$$

with respect to θ, subject to $\Sigma x=\zeta-m$. The latter equality does not determine x uniquely, but this does not matter for the following reason. Let ξ be any solution of $\Sigma\xi=\zeta-m$. Then $\Sigma(\xi-x)=0$, and hence, $\xi'\Sigma\xi=\xi'\Sigma(x+\xi-x)=\xi'\Sigma x=(x+\xi-x)'\Sigma x=x'\Sigma x$, so that the likelihood function is not affected.

III. Application to Partitioned Gaussian Random Vectors

We apply these results to the situation where the vector z is partitioned, and its mean and variance correspondingly, as

$$z=\begin{pmatrix}z_1\\z_2\end{pmatrix},\quad m=\begin{pmatrix}m_1\\m_2\end{pmatrix},\quad \Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}.$$

It is known that for Gaussian random vectors

$$E(z_2|z_1)=m_2+M(z_1-m_1),\qquad \operatorname{var}(z_2|z_1)=\Lambda_2$$

where M and Λ_2 are constant matrices that can be determined. It follows that

$$\begin{aligned}
\Sigma_{12}&=E(z_1-m_1)(z_2-m_2)'\\
&=E\{(z_1-m_1)E[(z_2-m_2)'|z_1]\}=\Sigma_{11}M'\\
\Sigma_{22}&=E(z_2-m_2)(z_2-m_2)'\\
&=E\{[z_2-m_2-(\hat{z}_2-m_2)+(\hat{z}_2-m_2)]\\
&\quad\cdot[z_2-m_2-(\hat{z}_2-m_2)+(\hat{z}_2-m_2)]'\}\\
&=\Lambda_2+E(\hat{z}_2-m_2)(\hat{z}_2-m_2)'\\
&=\Lambda_2+M\Lambda_{11}M'
\end{aligned}$$

where

$$\hat{z}_2=E(z_2|z_1).$$

Manuscript received September 14, 1978.

The author is with the Department of Applied Mathematics, Twente University of Technology, Enschede, The Netherlands.

0018-9286/79/0600-0496$00.75 ©1979 IEEE

Consequently,

$$\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{11}M'\\ M\Sigma_{11} & \Lambda_2+M\Sigma_{11}M'\end{pmatrix}.$$

With this result, and partitioning $x=\mathrm{col}(x_1,x_2)$, we can write

$$\begin{aligned}x'\Sigma x&=x_1'\Sigma_{11}x_1+2x_1'\Sigma_{11}M'x_2+x_2'(\Lambda_2+M\Sigma_{11}M')x_2\\&=x_2'\Lambda_2x_2+(x_1+M'x_2)'\Sigma_{11}(x_1+M'x_2)\\&=x_2'\Lambda_2x_2+\tilde{x}_1'\Sigma_{11}\tilde{x}_1\end{aligned}$$

where $\tilde{x}_1=x_1+M'x_2$. The equality constraint $\Sigma x=\zeta-m$ can now be written as

$$\Sigma_{11}x_1+\Sigma_{11}M'x_2=\zeta_1-m_1, \tag{3a}$$

$$M\Sigma_{11}x_1+(\Lambda_2+M\Sigma_{11}M')x_2=\zeta_2-m_2. \tag{3b}$$

Eliminating x_1 from (3b) with the aid of (3a) we obtain

$$\begin{aligned}\Lambda_2x_2&=\zeta_2-m_2-M(\Sigma_{11}x_1+\Sigma_{11}M'x_2)\\&=\zeta_2-m_2-M(\zeta_1-m_1)\\&=\zeta_2-\hat{\zeta}_2\end{aligned}$$

where $\hat{\zeta}_2=E(z_2|z_1=\zeta_1)$. For (3a) we may write $\Sigma_{11}\tilde{x}_1=\zeta_1-m_1$. Finally, we wish to find another expression for $\det(\Delta)$. Consider the following sequence of matrices, successively obtained by elementary transformations:

$$\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{11}M'\\ M\Sigma_{11} & \Lambda_2+M\Sigma_{11}M'\end{pmatrix}\rightarrow\begin{pmatrix}\Sigma_{11} & \Sigma_{11}M'\\ 0 & \Lambda_2\end{pmatrix}\rightarrow\begin{pmatrix}\Sigma_{11} & 0\\ 0 & \Lambda_2\end{pmatrix}.$$

Suppose that Σ_{11} and Λ_2 are, respectively, diagonalized as

$$\Sigma_{11}\sim\begin{pmatrix}\Delta_1 & 0\\ 0 & 0\end{pmatrix},\qquad \Lambda_2\sim\begin{pmatrix}\Delta_2 & 0\\ 0 & 0\end{pmatrix}$$

with Δ_1 and Δ_2 both positive definite. Then $\det(\Delta)=\det(\Delta_1)\det(\Delta_2)$. We thus conclude that application of the maximum likelihood method in the case of a partitioned random vector leads to the maximization of

$$\frac{1}{(2\pi)^{k/2}[\det(\Delta_1)\det(\Delta_2)]^{1/2}}\exp\left[-\frac{1}{2}(x_1'\Sigma_{11}x_1+x_2'\Lambda_2x_2)\right]$$

subject to

$$\Sigma_{11}x_1=\zeta_1-m_1$$

$$\Lambda_2x_2=\zeta_2-\hat{\zeta}_2$$

where we have omitted the tilde from x_2.

IV. Application to Linear Systems

We apply this result repeatedly to the random vector

$$y=\begin{bmatrix}y(0)\\ y(1)\\ \cdots\\ \cdots\\ y(N)\end{bmatrix}$$

with $y(k)$, $k=0,1,\cdots,N$, the solution of (1). Clearly, application of the maximum likelihood method now leads to maximization of the likelihood function

$$\frac{1}{(2\pi)^{k/2}\left[\prod_{i=0}^{N}\det(\Delta_i)\right]^{1/2}}\exp\left(-\frac{1}{2}\sum_{i=0}^{N}x_i'\Lambda_ix_i\right)$$

subject to

$$\Lambda_ix_i=y_i-\hat{y}_i,\qquad i=0,1,\cdots,N.$$

Here $y_0,y_1,\cdots,y_N$ are the observed realizations of the random variables $y(0),y(1),\cdots,y(N)$, $k=\dim(y)=(N+1)\dim[y(0)]$, while

$$\hat{y}_i=\begin{cases}E\{y(i)|y(0)=y_0,y(1)=y_1,\cdots,y(i-1)=y_{i-1}\} & \text{for } i=1,2,\cdots,N\\ E\{y(0)\} & \text{for } i=0\end{cases}$$

$$\Lambda_i=\begin{cases}E\{(y(i)-\hat{y}_i)(y(i)-\hat{y}_i)'|y(0)=y_0,y(1)=y_1,\cdots,y(i-1)=y_{i-1}\}\\ \quad\text{for } i=1,2,\cdots,N\\ \mathrm{var}\{y(0)\}\quad \text{for } i=0\end{cases}$$

Δ_i = nonsingular matrix occuring in the diagonalized form

$$\begin{pmatrix}\Delta_i & 0\\ 0 & 0\end{pmatrix}$$

of Λ_i. Using generalized inverses, we may equivalently write that maximum likelihood estimation amounts to maximization of the likelihood function

$$\frac{1}{(2\pi)^{k/2}\prod_{i=0}^{N}[\det(\Delta_i)]^{1/2}}\exp\left[-\frac{1}{2}\sum_{i=0}^{N}(y_i-\hat{y}_i)'\Lambda_i^{+}(y_i-\hat{y}_i)\right] \tag{4}$$

subject to

$$(I-\Lambda_i\Lambda_i^{+})(y_i-\hat{y}_i)=0,\qquad i=0,1,\cdots,N.$$

In case the matrices Λ_i, $i=0,1,\cdots,N$ are all nonsingular, this immediately specializes to the familiar result for maximum likelihood estimation (see, e.g., Sage and Melsa [1]).

The conditional expectations $\hat{y}_i$ and the conditional variances Λ_i may be obtained recursively with the aid of the Kalman filter as follows:

$$\hat{y}_i=C(i;\theta)\hat{x}_i$$

$$\Lambda_i=C(i;\theta)Q_i(\theta)C'(i;\theta)+V_{22}(i;\theta)$$

where

$$\begin{aligned}\hat{x}_{i+1}&=A(i;\theta)\hat{x}_i+B(i;\theta)u(i)+K_i(\theta)[y_i-C(i;\theta)\hat{x}_i]\\ K_i(\theta)&=[A(i;\theta)Q_i(\theta)C'(i;\theta)+V_{12}(i;\theta)]\\&\quad\cdot[V_{22}(i;\theta)+C(i;\theta)Q_i(\theta)C'(i;\theta)]^{+}\\ Q_{i+1}(\theta)&=[A(i;\theta)-K_i(\theta)C(i;\theta)]Q_i(\theta)A'(i;\theta)\\&\quad+V_{11}(i;\theta)-K_i(\theta)V_{12}'(i;\theta)\end{aligned}$$

with the initial conditions $\hat{x}_0=E\{x(0)\}$, $Q_0(\theta)=\mathrm{var}\{x(0)\}$.

V. Discussion

If $V_{22}(i;\theta)$ is positive definite for each i, clearly Λ_i is also always positive definite, and the likelihood function will never be singular. If $V_{22}(i;\theta)$ is singular for some or all i, the filtering problem is definitely singular. It is known from the literature (see, e.g., Tse and Athans [4]) that, in this case, the estimates $\hat{x}_i$ may be obtained with a recursive relation of a lower dimension than the Kalman filter given above. Even in the case of a singular filtering problem, however, the matrix Λ_i is not necessarily singular for all or some i. If, for example, V_{11} is nonsingular for each i (assuming for the moment that V_{12} vanishes and C has full rank for each i), Λ_i will always be positive definite. But even if V_{11} is singular, Λ_i may still remain nonsingular, depending on the structural properties of the system.

Let us contemplate for a moment the case that Λ_i is singular for some or all i. This problem has been studied for the time-invariant case where $V_{22}=0$ and $V_{12}=0$. It is known (Aoki [5]) that as i increases from 0, and given the right initial conditions, the range and null spaces of Q_i, and hence also those of Λ_i, vary monotonely, and become constant within a finite number of steps. Presumably, similar results hold in the case that V_{22} is nonzero but singular.

In case we are actually dealing with a singular maximum likelihood estimation problem, we have to maximize the likelihood function (4) subject to the constraints

$$(I-\Lambda_i\Lambda_i^+)(y_i-\hat{y}_i)=0, \qquad i=0,1,\cdots,N.$$

Suppose that the number of observations N is such that the number of equality constraints, as expressed by the last equation, exceeds the number of parameters in the vector θ, possibly even by quite a large number. In this case, the constraints necessarily are dependent. It is very well possible, however, that due to small numerical errors, the equations are, in fact, not dependent, and hence, a nonconsistent set of constraints is obtained. In this situation, the approach followed in this note breaks down. It then seems advisable to account for the small numerical errors by assuming some (small) measurement noise at the appropriate points, and thus slightly change V_{22} so that it becomes nonsingular. Solving the nonsingular maximum likelihood estimation problem thus obtained amounts to a penalty function approach for dealing with the equality constraints.

References

[1] A. P. Sage and J. L. Melsa, *System Identification.* New York: Academic, 1971.
[2] G. C. Goodwin and R. L. Payne, *Dynamic System Identification—Experiment Design and Data Analysis.* New York: Academic, 1977.
[3] B. Noble, *Applied Linear Algebra.* Englewood Cliffs, NJ: Prentice-Hall, 1969.
[4] E. Tse and M. Athans, "Optimal minimal-order observer-estimators for discrete linear time-varying systems," *IEEE Trans. Automat. Contr.*, vol. AC-15, pp. 416–426, Aug. 1970.
[5] M. Aoki, "On subspaces associated with partial reconstruction of state vectors, the structure algorithm, and the predictable directions of Riccati equations," *IEEE Trans. Automat. Contr.*, vol. AC-18, pp. 399–400, Aug. 1973.

Remarks on the Modulating Function Method for Impulse Response Identification

J. EISENFELD

***Abstract*—This note concerns the applications of the modulating function method to the input-output relation $y(t)=\int_0^t\phi(t-s)u(s)ds$. The problem is to determine the coefficients c_k in the differential equation $\sum_{k=0}^N c_k\phi^{(k)}(t)=0$ from integral transforms of $y(t)$ and $u(t)$ over a finite interval $[0,T]$. The main objective is to present a device which obtains the c_k's via the inversion of a system of N linear equations, whereas the straightforward approach leads to a system involving twice as many equations.**

I. Introduction

Among available methods for determination of coefficients entering linearly into a differential equation, Shinbrot's MFM (modulating function method) enjoys several advantages: linear, noniterative, no cutoff error [1], [2], and elimination of certain random and nonrandom errors [3]. We consider here the application of MFM to the input-output system

$$y(t)=\int_0^t\phi(t-s)u(s)\,ds, \qquad 0<t<T \tag{1}$$

where the impulse response $\phi(t)$ satisfies

$$c_N\phi^{(N)}+c_{N-1}\phi^{(N-1)}+\cdots+c_0\phi=0. \tag{2}$$

For convenience we set $c_N=1$ and "(k)" denotes the kth derivative. The problem is to determine the coefficients c_k from the waveforms $y(t)$ and $u(t)$ on $[0,T]$. The straightforward approach [2] converts the system (1)–(2) to a single differential equation $\sum_{k=0}^N c_k y^{(k)}=\sum_{j=0}^{N-1}a_j u^{(j)}$ which avoids the unknown ϕ. However, the parameters $a_j=\sum_{i=0}^{N-j-1}c_{i+j+1}\phi^{(i)}(0)$ are also unknown. Thus, $2N$ modulating functions are required to determine the $2N$ parameters, resulting in a system of $2N$ equations. Each modulating function $w_\nu(t)$ is required to satisfy the end-point conditions

$$\text{(i)} \qquad \omega_\nu^{(k)}(0)=0,$$
$$\text{(ii)} \qquad w_\nu^{(k)}(T)=0, \qquad k=0,1,\cdots,N-1. \tag{3}$$

Manuscript received December 28, 1978.
The author is with the Department of Medical Computer Science, University of Texas Health Science Center at Dallas, Dallas, TX, and the Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019.

II. Only N Equations are Required

If we set

$$w_\nu(t)=\int_0^{T-t}u(s)h_\nu(s+t)\,ds \tag{4}$$

where h_ν is chosen so as to satisfy

$$\text{(i)} \qquad \int_0^T h_\nu^{(k)}(t)u(t)\,dt=0,$$
$$\text{(ii)} \qquad h_\nu^{(k)}(T)=0, \qquad k=0,1,\cdots,N-1, \tag{5}$$

then $\omega_\nu(t)$ satisfies conditions (3) and, moreover,

$$(-1)^k s_{\nu k}\overset{\text{def}}{=}\int_0^T\phi(t)w_\nu^{(k)}(t)\,dt=\int_0^T y(t)h_\nu^{(k)}(t)\,dt. \tag{6}$$

Consequently, $s_{\nu k}$ may be computed from $y(t)$. Applying MFM to (2) yields

$$\sum_{k=0}^{N-1}s_{\nu k}c_k=-s_{\nu N}. \tag{7}$$

Thus, only N equations (corresponding to N method functions h_ν) are required to determine c_k. Observe that this approach does not require the differentiability of y and u, and thus it is more natural than the approach used in [2].

The functions w_ν, not used in the calculation, are merely a device to show the relationship of the method to MFM. Observe that the price paid for the reduction of equations is that end-point conditions 3(i) are traded for the orthogonality condition 5(i). These conditions may be satisfied as follows. Choose a set of $2N$ coordinate functions $\psi_i(t)$, each satisfying condition 5(ii), set $h_\nu=\sum_{i=1}^N b_{\nu i}\psi_{i+\nu-1}+\psi_{N+\nu}$, substitute into 5(i), and invert the resulting system of N equations for $b_{\nu i}$, $\nu=1,2,\cdots,N$. Notice that the choice of ψ_i determines the nature of the method (Fourier transforms, Laplace transforms, moments, etc.). See [1], [2], and [4] for discussions on the choice of method functions.

III. Reduction of Integrations

The calculation of each h_ν from the coordinate functions ψ_i requires the inversion of an $N\times N$ matrix whose elements $d_\nu(i,j)=\int_0^T\psi_{\nu+j-1}^{(i-1)}(t)u(t)\,dt$. Clearly, the number of integrations will be reduced if the coordinate functions are related by $\psi_{i+1}=(d/dt)\psi_i$, or equivalently, all the coordinate functions are defined in terms of a single prescribed function $\psi(t)$ by

$$\psi_i(t)=\psi^{(i-1)}(t), \qquad i=1,2,\cdots,N. \tag{10}$$

The penalty is that the function $\psi(t)$ will have to be chosen so that $\psi^{(i)}(T)=0$, $i=0,1,\cdots,2N-2$ which depresses the data more severely near $t=T$.

Similarly, the number of integrations required for the computation of the matrix $(s_{\nu k})$ is reduced if the h_ν are defined in terms of a single function $h(t)$ by

$$h_\nu(t)=-(-1)^\nu h^{(\nu-1)}(t), \qquad \nu=1,2,\cdots,N. \tag{11}$$